

# Scientific Experimentation for Cyber Security – Mission Impossible?

Xinming (Simon) Ou  
Kansas State University

1<sup>st</sup> Experimental Security Panoramas Workshop

# Who I am

- I do research on enterprise network security defense
  - Logic-based security analysis, attack graphs
  - Intrusion detection
  - Security metrics
- A common challenge I face everyday in my research
  - Evaluation of research methodologies

# Evaluation vs. Scientific Experiment

- What we call evaluation in Computer Science:
  - Run the tool on some loosely specified environment.
  - Get some numbers, draw diagrams, show that our method is cool.
  - How often do people try to repeat an experiment done by other people?
- What do people in other science disciplines do in experiments?

# Why experiments are even more difficult in cyber security

- The subject of experiments are often times humans.
  - E.g. effectiveness of IDS largely depends upon the intruder.
  - How to obtain an effective control is a big challenge.
  - For most researchers, we need data that serve as benchmarks for cyber-security measures' effectiveness.

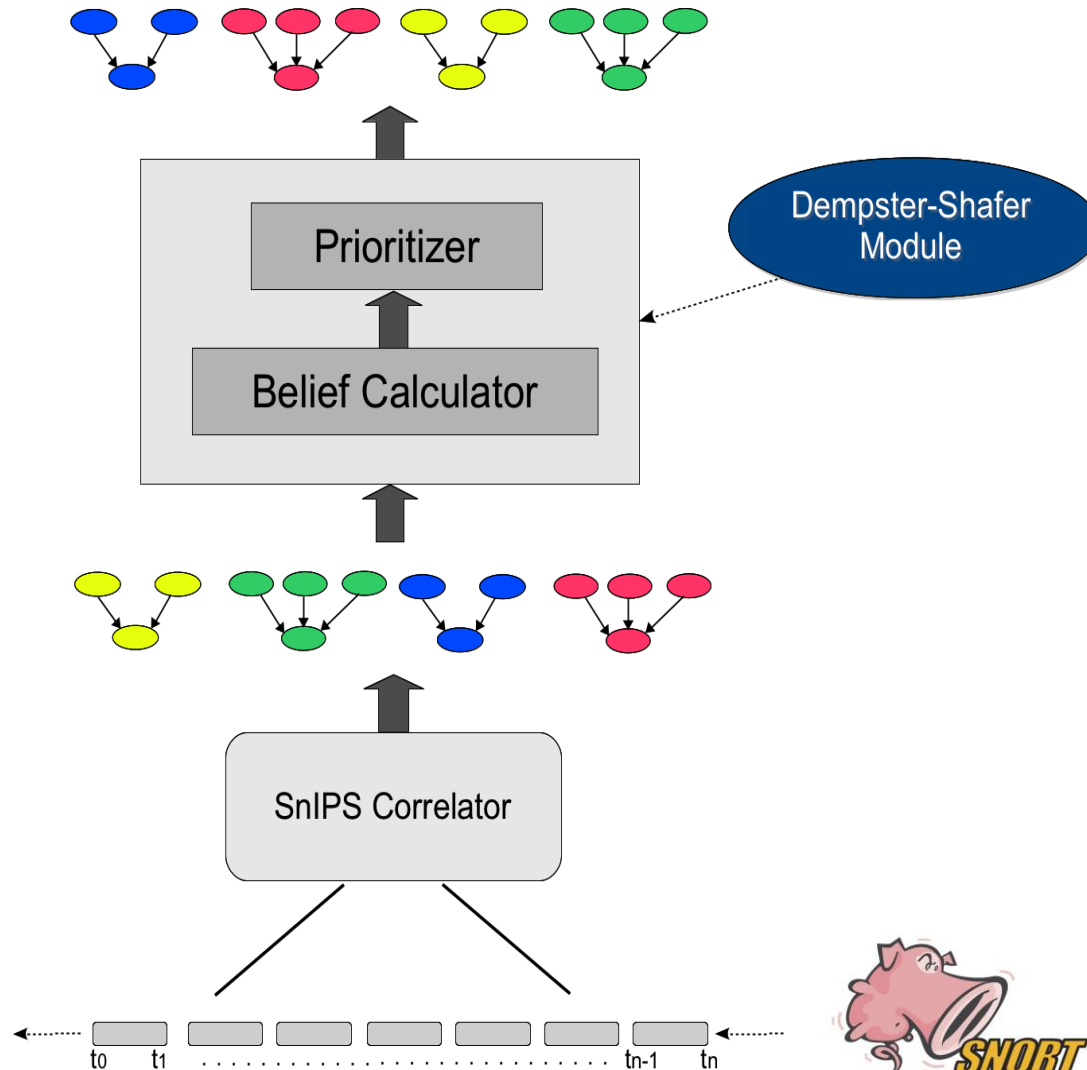
# But using data creates new problems

- Data lack ground truths, or need to be artificially created.
- Research methods can over-fit data.
- Famous example
  - MIT LL DARPA IDS Evaluation Datasets
  - [McHugh 2000], [Mahoney 2003]

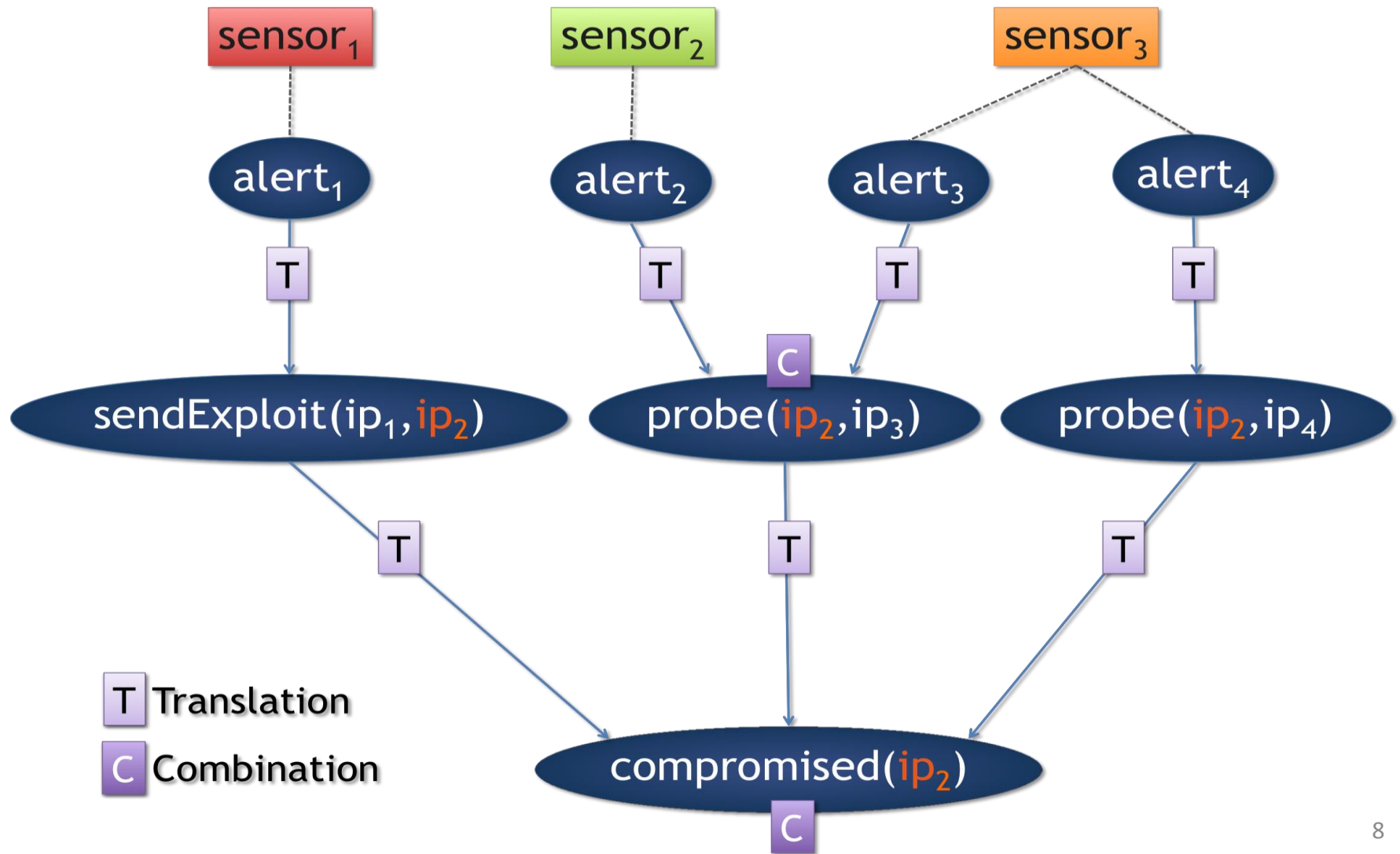
# Shall we stop doing the impossible?

- Risk for doing the experiments anyway
  - The validity of the result will be limited.
  - Could provide misleading conclusions.
- Risk for not doing
  - ???

# Experience: SnIPS IDS Analysis Tool

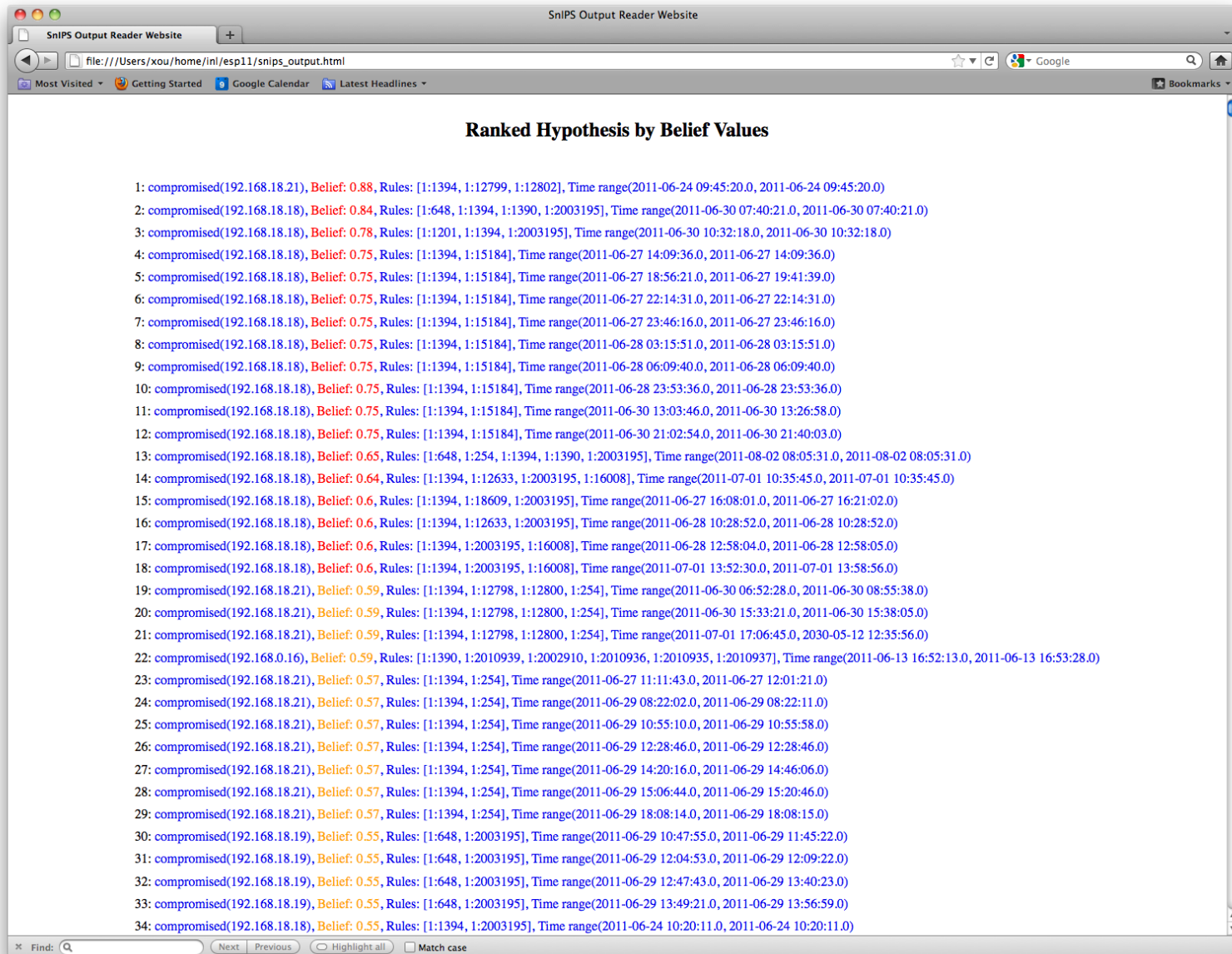


# Overview of the Dempster-Shafer theory calculation





# How can we know that D-S helps?



The screenshot shows a web browser window titled "SnIPS Output Reader Website". The address bar displays the file path: `file:///Users/xou/home/inl/esp11/snips_output.html`. The browser's toolbar includes "Most Visited", "Getting Started", "Google Calendar", "Latest Headlines", and "Bookmarks". The main content area is titled "Ranked Hypothesis by Belief Values" and lists 34 items, each representing a hypothesis with its belief value, rules, and time range. The items are numbered 1 through 34. The belief values range from 0.88 down to 0.55. The rules are listed in brackets, and the time ranges are in parentheses. The browser's status bar at the bottom shows "Find:" with a search icon, and buttons for "Next", "Previous", "Highlight all", and "Match case".

**Ranked Hypothesis by Belief Values**

- 1: compromised(192.168.18.21), **Belief: 0.88**, Rules: [1:1394, 1:12799, 1:12802], Time range(2011-06-24 09:45:20.0, 2011-06-24 09:45:20.0)
- 2: compromised(192.168.18.18), **Belief: 0.84**, Rules: [1:648, 1:1394, 1:1390, 1:2003195], Time range(2011-06-30 07:40:21.0, 2011-06-30 07:40:21.0)
- 3: compromised(192.168.18.18), **Belief: 0.78**, Rules: [1:1201, 1:1394, 1:2003195], Time range(2011-06-30 10:32:18.0, 2011-06-30 10:32:18.0)
- 4: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-27 14:09:36.0, 2011-06-27 14:09:36.0)
- 5: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-27 18:56:21.0, 2011-06-27 19:41:39.0)
- 6: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-27 22:14:31.0, 2011-06-27 22:14:31.0)
- 7: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-27 23:46:16.0, 2011-06-27 23:46:16.0)
- 8: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-28 03:15:51.0, 2011-06-28 03:15:51.0)
- 9: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-28 06:09:40.0, 2011-06-28 06:09:40.0)
- 10: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-28 23:53:36.0, 2011-06-28 23:53:36.0)
- 11: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-30 13:03:46.0, 2011-06-30 13:26:58.0)
- 12: compromised(192.168.18.18), **Belief: 0.75**, Rules: [1:1394, 1:15184], Time range(2011-06-30 21:02:54.0, 2011-06-30 21:40:03.0)
- 13: compromised(192.168.18.18), **Belief: 0.65**, Rules: [1:648, 1:254, 1:1394, 1:1390, 1:2003195], Time range(2011-08-02 08:05:31.0, 2011-08-02 08:05:31.0)
- 14: compromised(192.168.18.18), **Belief: 0.64**, Rules: [1:1394, 1:12633, 1:2003195, 1:16008], Time range(2011-07-01 10:35:45.0, 2011-07-01 10:35:45.0)
- 15: compromised(192.168.18.18), **Belief: 0.6**, Rules: [1:1394, 1:18609, 1:2003195], Time range(2011-06-27 16:08:01.0, 2011-06-27 16:21:02.0)
- 16: compromised(192.168.18.18), **Belief: 0.6**, Rules: [1:1394, 1:12633, 1:2003195], Time range(2011-06-28 10:28:52.0, 2011-06-28 10:28:52.0)
- 17: compromised(192.168.18.18), **Belief: 0.6**, Rules: [1:1394, 1:2003195, 1:16008], Time range(2011-06-28 12:58:04.0, 2011-06-28 12:58:05.0)
- 18: compromised(192.168.18.18), **Belief: 0.6**, Rules: [1:1394, 1:2003195, 1:16008], Time range(2011-07-01 13:52:30.0, 2011-07-01 13:58:56.0)
- 19: compromised(192.168.18.21), **Belief: 0.59**, Rules: [1:1394, 1:12798, 1:12800, 1:254], Time range(2011-06-30 06:52:28.0, 2011-06-30 08:55:38.0)
- 20: compromised(192.168.18.21), **Belief: 0.59**, Rules: [1:1394, 1:12798, 1:12800, 1:254], Time range(2011-06-30 15:33:21.0, 2011-06-30 15:38:05.0)
- 21: compromised(192.168.18.21), **Belief: 0.59**, Rules: [1:1394, 1:12798, 1:12800, 1:254], Time range(2011-07-01 17:06:45.0, 2030-05-12 12:35:56.0)
- 22: compromised(192.168.0.16), **Belief: 0.59**, Rules: [1:1390, 1:2010939, 1:2002910, 1:2010936, 1:2010935, 1:2010937], Time range(2011-06-13 16:52:13.0, 2011-06-13 16:53:28.0)
- 23: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-27 11:11:43.0, 2011-06-27 12:01:21.0)
- 24: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 08:22:02.0, 2011-06-29 08:22:11.0)
- 25: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 10:55:10.0, 2011-06-29 10:55:58.0)
- 26: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 12:28:46.0, 2011-06-29 12:28:46.0)
- 27: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 14:20:16.0, 2011-06-29 14:46:06.0)
- 28: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 15:06:44.0, 2011-06-29 15:20:46.0)
- 29: compromised(192.168.18.21), **Belief: 0.57**, Rules: [1:1394, 1:254], Time range(2011-06-29 18:08:14.0, 2011-06-29 18:08:15.0)
- 30: compromised(192.168.18.19), **Belief: 0.55**, Rules: [1:648, 1:2003195], Time range(2011-06-29 10:47:55.0, 2011-06-29 11:45:22.0)
- 31: compromised(192.168.18.19), **Belief: 0.55**, Rules: [1:648, 1:2003195], Time range(2011-06-29 12:04:53.0, 2011-06-29 12:09:22.0)
- 32: compromised(192.168.18.19), **Belief: 0.55**, Rules: [1:648, 1:2003195], Time range(2011-06-29 12:47:43.0, 2011-06-29 13:40:23.0)
- 33: compromised(192.168.18.19), **Belief: 0.55**, Rules: [1:648, 1:2003195], Time range(2011-06-29 13:49:21.0, 2011-06-29 13:56:59.0)
- 34: compromised(192.168.18.18), **Belief: 0.55**, Rules: [1:1394, 1:2003195], Time range(2011-06-24 10:20:11.0, 2011-06-24 10:20:11.0)

# Experiment Strategy

- We need data with ground truth
  - Use production system, with assistance from system administrators
    - Highly labor intensive
    - Hard to justify the result
  - Decided to use MIT LL DARPA dataset
    - It has many limitations.
    - It has been harshly criticized in the literature.
    - But it is the only publicly available IDS dataset with ground truth.

# Avoid the Pitfalls in the LL Dataset

- Artificially generated attack data can easily lead to over-fitting
  - By just looking at the TTL field of an IP packet one would be able to tell attack and non-attack packets apart [Mahoney 2003].
  - This can easily lead to over-fitting, especially for learning-based methods.
- Do not train the model on the dataset

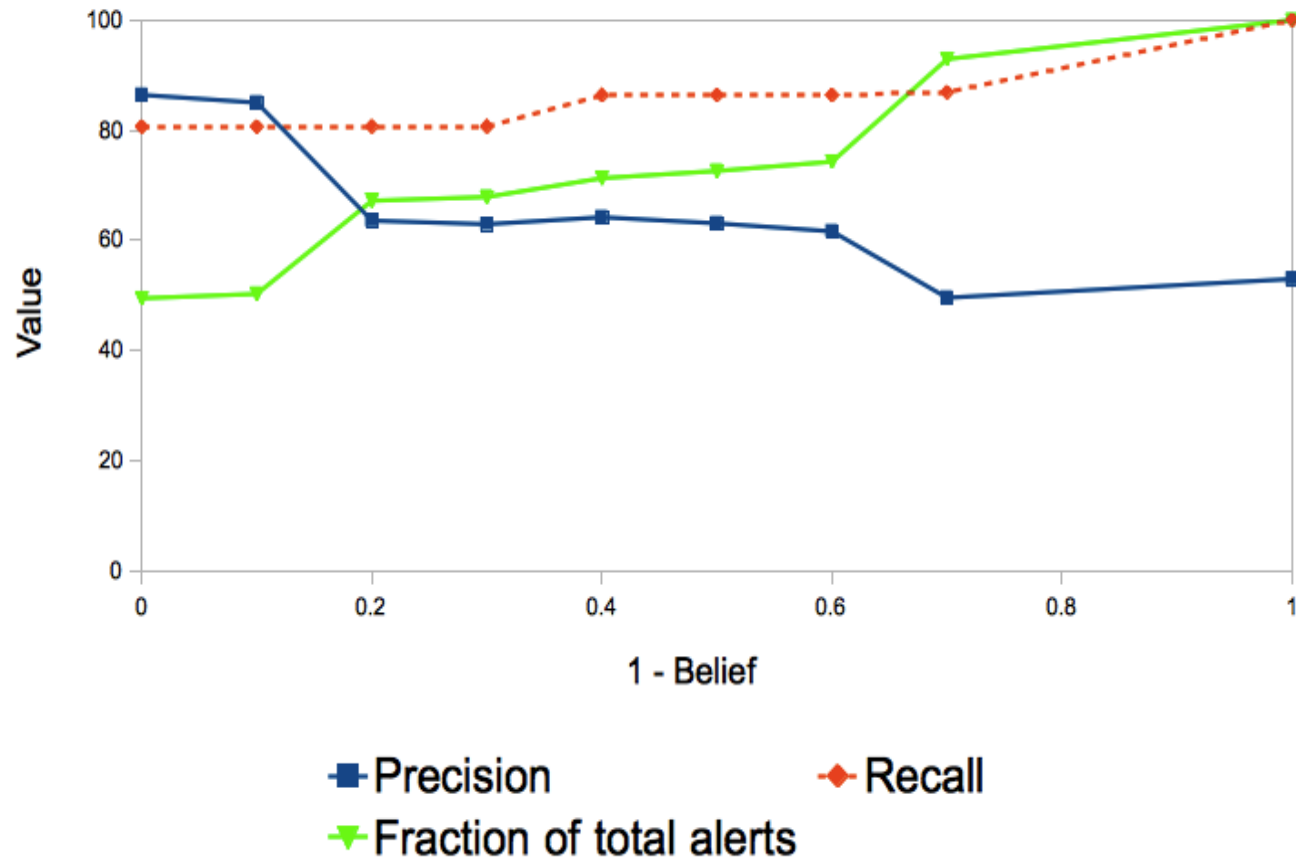
# Avoid the Pitfalls in the LL Dataset

- Background traffic is low [McHugh 2000]
  - The prior probability of an event being true attack is much higher than a production system
    - About half the Snort alerts are true alerts
  - This makes it easier to have good detection rate and false positive rate.
- Do not claim performance on the absolute false positive and negative rates

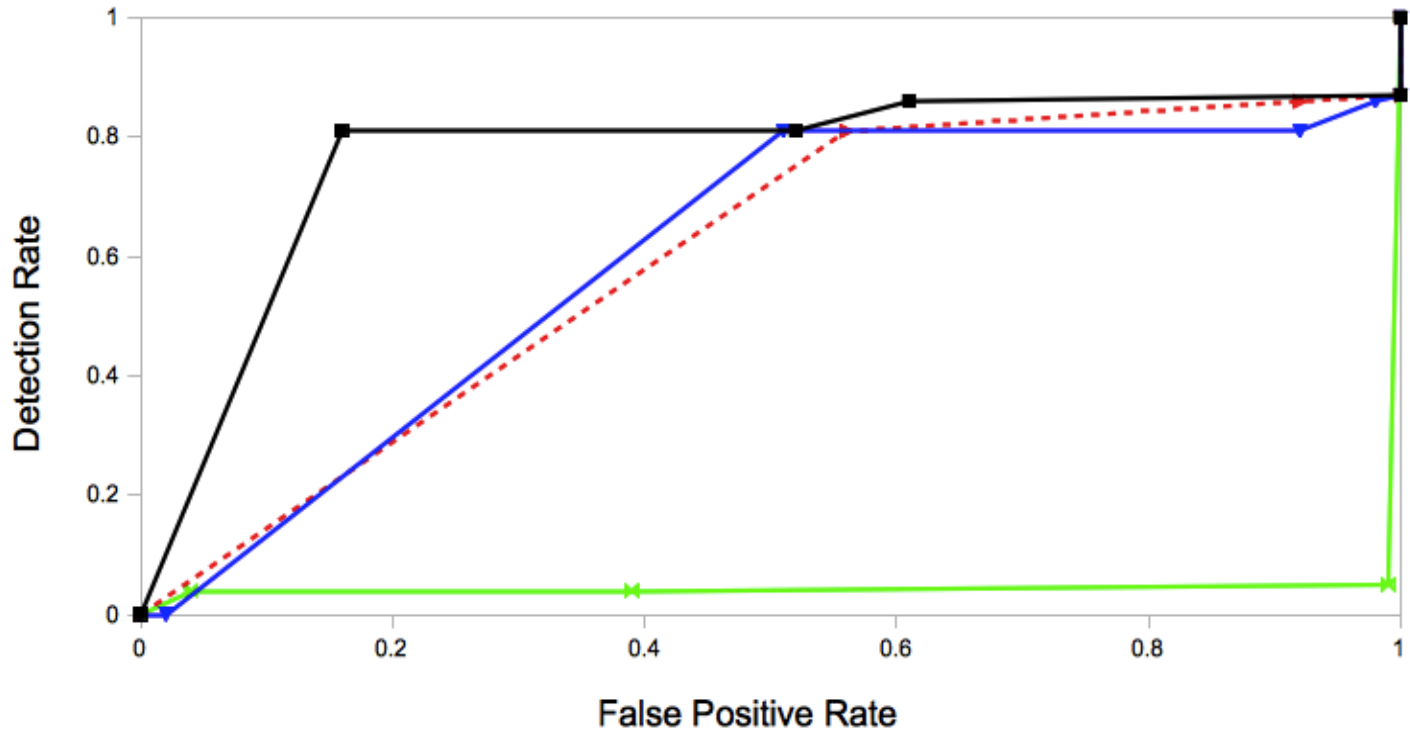
# Then what do we evaluate?

- Will the ranking provided by Dempster-Shafer belief calculation indeed help in prioritizing IDS alerts?
- Is it really D-S that helps?

# Prioritization Effect

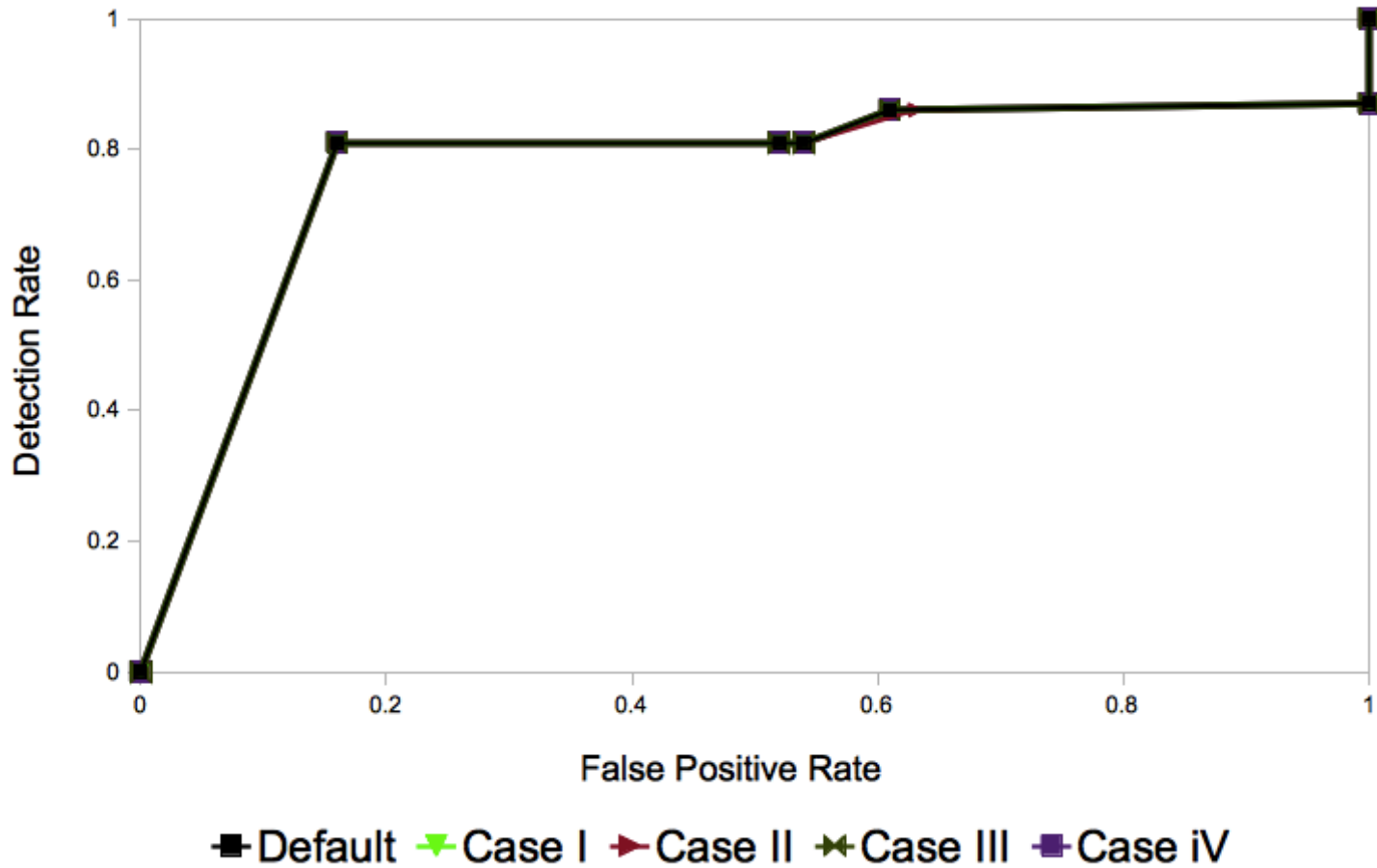


# ROC Curve



■ Customized DS      ▼ Standard DS  
▶ Max. Mode in the Graph      x Sensor Quality Metrics

# Sensitivity Analysis





# Lessons learned from the experience

- Even flawed data could produce insights into a security method's effectiveness.
  - We shall not easily write-off datasets like DARPA IDS evaluation data.
  - But the experiments must be designed carefully to avoid the flaws to the maximum degree possible.
- We need more (flawed) data like this!

# Discussion

Questions?